Journal of Biometrics & Biostatistics
ISSN: 2155-6180

Breheny et al., J Biom Biostat 2012, 3:7
DOI: 10.4172/2155-6180.1000e118

Editorial Open Access

Statistical Challenges and Opportunities in Copy Number Variant Association Studies

Patrick Breheny[1,2]*, Yinglei Li[2] and Richard Charnigo[1,2]

[1]Department of Biostatistics, University of Kentucky, 725 Rose Street, Lexington KY, USA
[2]Department of Statistics, University of Kentucky, 725 Rose Street, Lexington KY, USA

According to classical genetics, humans have two copies of each region of DNA. During the past decade, however, a large body of research has emerged demonstrating that this is something of an oversimplification [1-3]. Even phenotypically normal individuals have many stretches of their genome in which more or fewer than two copies are found – these stretches have been estimated to constitute roughly 5% of the entire genome [4]. Such genetic variations are referred to as Copy Number Variants (CNVs).

In the wake of the Human Genome Project, a tremendous effort has been spent on understanding the genetic basis of human variation and disease through Genome-Wide Association (GWA) studies. The vast majority of this effort has focused on one-base-pair differences between individuals, termed Single Nucleotide Polymorphisms (SNPs). As the research on copy-number variation demonstrates, however, SNPs represent only one type of genetic variation.

One effort to quantify the relative contributions of SNPs and CNVs on gene expression estimated that SNPs were responsible for 84% of the explainable variation, while CNVs were responsible for 17%, with only 1% resulting from overlapping effects [5]. This 17% represents a sizable degree of genetic variation that has been understudied. Furthermore, some have argued that CNVs are more likely, a priori, to play a role in common diseases because, given that they represent a more subtle, quantitative genetic variation, they are less likely to have been selected out of the population by evolutionary pressures [6]. Indeed, CNVs have been linked to a number of diseases such as Crohn's disease, psoriasis, and autism [7-9].

Fortunately, CNV information can be mined from existing data collected by GWA studies, thereby avoiding the considerable costs of carrying out new studies. One of the limiting factors-perhaps the limiting factor-in carrying out genome-wide CNV association studies, however, is a challenge of analyzing the data. While methods to determine the locations in which an individual has gained or lost copies of genetic material are fairly well-developed [10,11], methods for integrating these CNV calls into an association study are "still in [their] infancy" [12]. Relative to that of CNVs, genome-wide analysis of SNPs is straightforward: at every genetic marker, each individual is genotyped (AA, AB, or BB) and an association test is carried out. An adjustment for multiple comparisons then preserves the overall type I error rate.

A similar analytic strategy does not readily apply to CNV association studies, for two primary reasons. First, the uncertainty in a CNV call is much greater than that in a SNP call. For each type of calling, the goal is to classify a sample into one of three groups (AA/AB/BB for a SNP, gain/loss/neutral for a CNV) based on probe intensity measurements. However, for a SNP, one obtains a two-dimensional measure (intensities for both the A and B probes); for CNVs, one obtains only a one-dimensional measure (total intensity). Consequently, there is a much greater separation between classes for SNPs, and more extensive misclassification in CNV genotyping.

The second reason is that, unlike SNPs, CNVs span multiple markers and introduce an added complexity: that of estimating the boundaries of the CNV. These two features of CNV data are illustrated in the left panel of figure 1, which comes from an analysis of real data described in Breheny et al. [13]. As noted earlier, CNV calling is based on a one-dimensional measure of intensity; the gray region was determined to have a loss of copy, leading to lower intensity throughout that region. As figure 1 indicates, there is no clear separation between intensities originating from the white (neutral) and gray (loss) regions. Furthermore, the precise boundaries of the CNV are not obvious.

Each of these two features of CNV data complicates association testing. Ignoring misclassification error may considerably diminish the power of a test [14], while the imprecise estimation of boundaries makes it difficult to determine whether two partially overlapping CNVs represent the same genetic variation. Reasonable decision rules for two overlapping CNVs may be proposed; however, even with as few as three CNVs, patterns may arise for which there is no unambiguous resolution. For example, consider a scenario with three CNVs: A, B, and C. Suppose A has 50% overlap with B, B has 50% overlap with C, but A and C have no overlap. How many association tests should one carry out? A variety of ad-hoc rules have been proposed to address this scenario, but one can easily imagine how intractable the problem may become with, say, 25 partially overlapping CNVs. Dealing with partial overlap is both burdensome in practice and likely to be statistically inefficient.

To avoid these complications, one may avoid CNV calling altogether and carry out marker-level testing (as opposed to the previous approach, which we refer to as variant-level testing). Marker-level tests can be simple, such as carrying out t-tests of CNV intensities between cases and controls, or more complex, involving mixture models to incorporate uncertainty in copy number [15]. However, due to the noise in intensity measurements, single-marker tests tend to have low power and require very large sample sizes (n>4,000 for the study in Barnes et al. [15]).

An intriguing possibility is to supplement the power of single-marker tests by pooling results from neighboring tests. This idea is illustrated in the right panel of figure 1. If a CNV spanning multiple marker is present and associated with the phenotypic outcome of interest, this will induce marker-level associations spanning the genomic region covered by the CNV. Although no single p-value in

***Corresponding author:** Patrick Breheny, Department of Biostatistics, University of Kentucky, 725 Rose Street, Lexington KY, USA, E-mail: patrick.breheny@uky.edu

Received November 17, 2012; **Accepted** November 19, 2012; **Published** November 26, 2012

Citation: Breheny P, Li Y, Charnigo R (2012) Statistical Challenges and Opportunities in Copy Number Variant Association Studies. J Biom Biostat 3:e118. doi:10.4172/2155-6180.1000e118

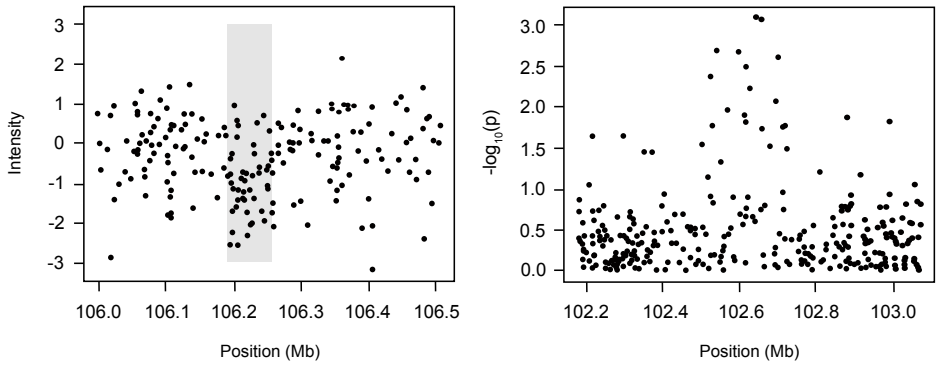

Figure 1: Left: Illustration of CNV calling. Vertical axis is total intensity, normalized so that 0 corresponds to two copies. The gray shading indicates a region identified as a CNV. Right: Illustration of marker-level testing. The *p*-values from marker-level tests are plotted on the negative log scale.

figure 1 is particularly small, the fact that so many low p-values are present in a single cluster is suggestive of a CNV that is associated with the outcome. Breheny et al. [13] proposed this idea and presented evidence suggesting that aggregation of marker-level tests could prove more powerful than both single-marker testing and variant-level testing. A more careful examination of marker-level test aggregation was conducted in Li and Breheny [16], which demonstrated that proper inference under aggregation is not trivial. The null distribution for any quantity which aggregates marker-level tests is complicated by the fact that a CNV can span multiple markers, thereby introducing local correlations among the test results even under the null hypothesis. This violates exchangeability among the marker-level tests and invalidates simple approaches to deriving a null distribution. In Li and Breheny [16], the authors proposed a permutation-based approach for estimating the empirical null distribution in a way that preserves the local correlations among nearby tests. In addition, they proved that their approach maintains the correct family-wise error rate for a genome-wide analysis. One downside of their approach, of course, is its computationally intensive nature, which would impede the use of sophisticated marker-level tests such as those in Barnes et al. [15]. Whether there exist other, less intensive methods for aggregating test results across markers remain to be seen.

There are interesting possibilities for improving variant-level tests as well, based on the idea of joint CNV calling. Rather than calling CNVs separately for each sample, several authors [17-19] have recently proposed methods for jointly calling common CNVs across multiple samples, potentially eliminating the partial overlap issue discussed earlier. These methods are still new, however, and the feasibility of extending them to CNV association studies on the genome-wide level has not yet been investigated.

Our focus in this editorial has been on analytical approaches for incorporating information across markers and across samples when performing association tests. We do not wish to downplay other important statistical issues in CNV association studies, such as normalization of the data, proper experimental design, and controlling for confounding factors. Rather, we hope to have highlighted some interesting features of CNV data, limitations of existing approaches, possible avenues for improvement, and open statistical questions surrounding this important area of scientific inquiry.

References

1. Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK , et al. (2004) Detection of large-scale variation in the human genome. Nat Genet 36: 949-951.
2. Sebat J, Lakshmi B, Troge J, Alexander J, Young J, et al. (2004) Large-scale copy number polymorphism in the human genome. Science 305: 525-528.
3. Tuzun E, Sharp AJ, Bailey JA, Kaul R, Morrison VA, et al. (2005) Fine-scale structural variation of the human genome. Nat Genet 37: 727-732.
4. Sharp AJ, Cheng Z, Eichler EE (2006) Structural variation of the human genome. Annu Rev Genomics Hum Genet 7: 407-442.
5. Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, et al. (2007) Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science 315: 848-853.
6. Plagnol V, Clayton D (2011) Copy number variant association studies. Analysis of Complex Disease Association Studies: A Practical Guide.
7. Fellermann K, Stange DE, Schaeffeler E, Schmalzl H, Wehkamp J, et al. (2006) A chromosome 8 gene-cluster polymorphism with low human beta-defensin 2 gene copy number predisposes to Crohn disease of the colon. Am J Hum Genet 79: 439-448.
8. Sebat J, Lakshmi B, Malhotra D, Troge J, Lese-Martin C, et al. (2007) Strong association of de novo copy number mutations with autism. Science 316: 445-449.
9. Hollox EJ , Huffmeier U, Zeeuwen PL, Palla R, Lascorz J, et al. (2008) Psoriasis is associated with increased beta-defensin genomic copy number. Nat Genet 40: 23-25.
10. Olshen AB, Venkatraman E, Lucito R, Wigler M (2004) Circular binary segmentation for the analysis of array-based DNA copy number data. Biostatistics 5: 557-572.
11. Wang K , Li M, Hadley D, Liu R, Glessner J, et al. (2007) PennCNV: An integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. Genome Res 17: 1665-1674.
12. Ionita-Laza I, Rogers AJ, Lange C, Raby BA, Lee C (2009) Genetic association analysis of copy-number variation (CNV) in human disease pathogenesis. Genomics 93: 22-26.
13. Breheny P, Chalise P, Batzler A, Wang L, Fridley BL (2012) Genetic Association Studies of Copy-Number Variation: Should Assignment of Copy Number States Precede Testing? PLoS ONE 7: e34262.
14. McCarroll SA, Altshuler DM (2007) Copy-number variation and association studies of human disease. Nat Genet 39: S37-S42.
15. Barnes C, Plagnol V, Fitzgerald T, Redon R, Marchini J, et al. (2008) A robust statistical method for case-control association testing with copy number variation. Nature Genetics 40: 1245-1252.
16. Li Y, Breheny P (2012) Kernel-based aggregation of marker-level genetic association tests involving copy-number variation. Technical Report, University of Kentucky, Department of Statistics.
17. Efron B, Zhang NR (2011) False discovery rates and copy number variation. Biometrika 98: 251-271.

18. Nowak G, Hastie T, Pollack JR, Tibshirani R (2011) A fused lasso latent feature model for analyzing multi-sample a CGH data. Biostatistics 12: 776-791.

19. Zhang Z, Lange K, Sabatti C (2012) Reconstructing DNA copy number by joint segmentation of multiple sequences.